Research Report

Sample-Size Requirements for Automated Essay Scoring

Shelby J. Haberman

Sandip Sinharay

June 2008

ETS RR-08-32

Sample-Size Requirements for Automated Essay Scoring

Shelby J. Haberman and Sandip Sinharay

ETS, Princeton, NJ

June 2008

Abstract

Sample-size requirements were considered for automated essay scoring in cases in which the automated essay score estimates the score provided by a human rater. Analysis considered both cases in which an essay prompt is examined in isolation and those in which a family of essay prompts is studied. In typical cases in which content analysis is not employed and in which the only object is to score individual essays to provide feedback to the examinee, it appears that several hundred essays are sufficient. For application of one model to a family of essays, fewer than 100 essays per prompt may often be adequate. The cumulative logit model was explored as a possible replacement of the linear regression model usually employed in automated essay scoring; the cumulative logit model performed somewhat better than did the linear regression model.

Key words: Cross-validation, PRESS, residual, regression

Acknowledgments

We thank Dan Eignor, David Williamson, Jiahe Qian, and Frank Rijmen for their helpful comments; Cathy Trapani and Sailesh Vezzu for their help with the data used in this work; and Kim Fryer for her help with proofreading.

Automated essay-scoring programs such as e-rater® (Attali, Burstein, & Andreyev, 2003; Burstein, Chodorow, & Leacock, 2004) use samples of essays that have been scored by human raters in order to estimate prediction equations in which the dependent variable is a human essay score obtained by an examinee on an essay prompt and the independent variables are computer-generated essay features for that essay. The prediction equations may be applied to predict human scores on essays in which the computer-generated features are available but no human scores exist. Prediction may be based on linear regression, as is currently the case with e-rater, or may be based on techniques such as cumulative logit analysis (Haberman, 2006).

A basic question to consider with automated essay scoring is the sample size required for satisfactory estimation of the prediction equations. Currently, the e-rater V.2 software (Attali & Burstein, 2006) requires a sample size of 500 to build the regression model and another sample of 500 to cross-validate the regression model. Interestingly, after cross-validation, the model is not re-estimated. The processing waits until the pre-assigned sample size is reached. The sample-size selection problem may be addressed by use of the cross-validation techniques described in Haberman (2006); however, some variations on the approach are appropriate when a family of essay prompts is available rather than a single essay prompt. In addition, some added variations are reasonable to consider to avoid problems with outliers.

Some of the methodologies discussed in this paper were discussed in Haberman (2006), and this paper follows Haberman (2006) in its use of mean-squared error to assess prediction quality. This emphasis on mean-squared error reflects common statistical practice with linear estimation and reflects the common practice in testing of adding item scores in assessments. However, this paper expands the methods of Haberman (2006), provides applications of the methods to a wider number of prompts, and discusses the application of the methods discussed in Haberman (2006) to a family of essay prompts.

Section 1. describes the data used in this work. Section 2. considers some basic screens for outliers in essay features that may cause distortions in analysis. Section 3. describes use of deleted residuals to assess prediction accuracy and to estimate the loss of precision expected for a given sample size. In this section, applications are made to four prompt families for several different approaches based on linear regression. In section 4., alternative analysis based on cumulative logit models is considered. Some practical conclusions are provided in section 5.. Familiarity with standard works on regression analysis (Draper & Smith, 1998; Weisberg, 1985) and classical test theory (Lord & Novick, 1968) is helpful in reading this report.

1. Data

The data used in the analysis are four groups of prompts. Group 1 consists of 74 prompts associated with a particular licensure test, Group 2 includes 14 prompts from a discontinued test of knowledge of English for examinees whose native language is not English, Group 3 comprises 26 prompts for a graduate admissions examination, and Group 4 uses 16 practice prompts for a test of knowledge of English administered to examinees whose native language is not English. For each prompt, about 500 essays are available. In Group 1, human scores are on a 4-point scale, and only a small number of essays are double-scored. Groups 2 and 3 use double-scoring and have a 6-point scale, while Group 4 has double-scoring and a 5-point scale. In all cases, 1 is the lowest score for a valid essay. As a consequence, no ratings out of the range of a valid essay were used in this study.

In the analysis in this report, predictors used were logarithm of number of discourse elements (logdtu), logarithm of average number of words per discourse element (logdta), minus the square root of the number of grammatical errors detected per word (nsqg), minus the square root of the number of mechanics errors detected per word (nsqm), minus the square root of the number of usage errors detected per word (nsqu), minus the square root of the number of style errors detected per word (nsqstyle), minus the median Standard Frequency Index of words in the essay for which the index can be evaluated (nwfmedian), and average word length (wordln2). Features were based on those used in e-rater Version 7.2 for models without content vector analysis. The signs were selected so that the normal sign of the corresponding regression coefficient would be positive.

Table 1.
Frequency of Essays With Fewer Than 25 Words

Group	Number in group	Number very short in group	Fraction very short in group
1	37,000	346	0.0094
2	6,384	0	0.0000
3	12,251	117	0.0096
4	8,036	41	0.0051

2. Outlier Screens

It is prudent in any analysis of essays to exclude submissions that are too short to be meaningful and those with feature outliers that suggest major typing errors. As in Haberman (2006), the rule was adopted that any essay considered must have at least 25 words. This restriction removes many cases in which essay features exhibit unusual behavior. An added rule was adopted that no essay in which the average word length exceeded 7.5 characters be considered. The issue here is that such a case is likely to involve an error by the writer in using the keyboard. For example, it may occur if the space bar is not used properly.

The restriction on number of words involves an appreciable number of essays, as is evident in Table 1.

Except in Group 1, all human scores for essays with no more than 25 words received the lowest possible score. In Group 1, 8 cases of 346 received human scores of 2 rather than 1. In contrast, for essays with at least 25 words in Group 1, only about 13% received scores of 1.

The restriction on average word length affected very few essays. In the case of Group 1, two such essays arose, and one essay appeared in Group 3. All received scores of 1 and also had no more than 25 words.

An alternative check on outliers involved an examination of standardized values of variables that exceeded 4 in magnitude. The standardization was conducted for all essays for a prompt after removing those essays with no more than 25 words. In examining results, it is helpful to note that a feature with a normal distribution would yield standardized values of magnitude four or greater with probability 0.00006. Observed rates were somewhat higher for most features, as is evident in Table 2; however, the results were not unusual for relatively conventional distributions. For example, under a logistic distribution, the probability of a standardized value of magnitude at least 4 is 0.0014. For an exponential distribution, this probability is 0.0183. The results of regression analysis described in this paper provided no compelling reason to remove outliers other than essays with fewer than 25 words or an average word length above 7.5. In general, outliers must be rather extreme before they have significant impact on the analysis in this report. This impact normally will be evident through the analysis of variance inflation in Section 3.. Outliers are a concern if the estimated variance inflation is much higher than usually encountered for a given sample size and number of predictors; however, as already indicated, no case requiring consideration of outliers was encountered that is not associated with average word length or number of words in the essay.

3. Sample-Size Determination for the Linear Regression Model

In the case of linear regression, deleted residuals provide a basic method for assessing the accuracy of predicting results of human scoring when applied to data not used to estimate regression parameters (Haberman, 2006). In essence, deleted residuals provide an approach to cross-validation that requires only minimal computations and provides much higher accuracy than primitive approaches in which half a sample is used to construct a model and half a sample is used to examine prediction quality. We will first lay out the statistical model for essay scoring and provide expressions for several mean-squared errors, proportional reductions, and relative increases in mean-squared error that are crucial in sample-size determination. We will then discuss how to estimate the above mentioned quantities using deleted residuals.

Table 2.
Frequency of Outliers

Group	Variable	Number in group	Number of outliers in group	Fraction of outliers in group
1	logdta	36,697	43	0.0012
1	logdtu	36,697	0	0.0000
1	nwfmedian	36,697	104	0.0028
1	nsqg	36,697	86	0.0023
1	nsqm	36,697	196	0.0053
1	nsqu	36,697	104	0.0028
1	nsqstyle	36,697	3	0.0001
1	wordln2	36,697	31	0.0008
2	logdta	6,384	14	0.0024
2	logdtu	6,384	0	0.0000
2	nwfmedian	6,384	9	0.0014
2	nsqg	6,384	15	0.0024
2	nsqm	6,384	14	0.0022
2	nsqu	6,384	16	0.0025
2	nsqstyle	6,384	0	0.0000
2	wordln2	6,384	2	0.0003
3	logdta	12,143	18	0.0015
3	logdtu	12,143	5	0.0004
3	nwfmedian	12,143	3	0.0002
3	nsqg	12,143	26	0.0021
3	nsqm	12,143	28	0.0023
3	nsqu	12,143	17	0.0014
3	nsqstyle	12,143	0	0.0000
3	wordln2	12,143	5	0.0004
4	logdta	7,997	6	0.0008
4	logdtu	7,997	3	0.0004
4	nwfmedian	7,997	17	0.0021
4	nsqg	7,997	9	0.0011
4	nsqm	7,997	9	0.0011
4	nsqu	7,997	7	0.0009
4	nsqstyle	7,997	0	0.0000
4	wordln2	7,997	6	0.0008

3.1 The Statistical Model

Consider a random sample of essays i, $0 \leq i \leq n$. For some positive integer $q < n - 1$, for essay i, let Y_{ij} be the holistic score provided by reader j, $1 \leq j \leq m_i$, $m_i \geq 1$, and let $\mathbf{X}_i$ be a q-dimensional vector with coordinates X_{ik}, $1 \leq k \leq q$, that are numerical features of the essay that have been generated by computer processing. For example, X_{i1} might be the observed logdta for essay i, and q might be 8. Let $\bar{Y}_i$ be the average of the Y_{ij}, $1 \leq j \leq m_i$. Assume that the holistic scores are integers from 1 to G for an integer $G \geq 2$, and assume that the X_{ik} all have finite fourth moments and that the covariance matrix $\text{Cov}(\mathbf{X})$ of $\mathbf{X}_i$ is positive-definite. In the simplest cases, m_i is a fixed value m for all essays. More generally, the m_i are independent random variables that are independent of the X_{ik} and Y_{ij} and each $m_i \leq m$ for some given integer $m \geq 1$. In typical applications, details of the rating process are quite limited, so that it is appropriate to assume that independent and identically distributed random variables T_i, the true essay scores, exist such that

$$Y_{ij} = T_i + e_{ij},$$

T_i has positive finite variance σ_T^2, and the scoring errors e_{ij} are all uncorrelated, have mean 0, have common positive variances σ^2, and are uncorrelated with T_i and the X_{ik}'s.

The assumptions on the errors can be violated if the same rater scores many essays from many examinees and if the conditional distribution of e_{ij} depends on the specific rater who provides score j for essay i. Because virtually all data involve far fewer raters than examinees, the assumptions on the e_{ij} are not entirely innocuous. Without data in which raters are identified, it is impossible to investigate the implications of assignment of the same rater to many essays. It appears that the methods used in this report can still be used if the probability that two essays receive the same rater is the same for all pairs of essays.

A further possible violation of assumptions arises when essay features X_{ik} are used that depend on properties of essays other than essay i. This issue arises in practice in e-rater when content vector analysis is considered. The approach in this report does not apply to features associated with content vector analysis (Attali & Burstein, 2006).

Consider use of ordinary least squares with essays i from 1 to n to estimate the coefficients α and β_k, $1 \leq k \leq q$, that minimize the mean-squared error

$$\sigma_d^2 = E(d_i^2) \tag{1}$$

for

$$d_i = \bar{Y}_i - T_i^*,$$

where

$$T_i^* = \alpha + \sum_{k=1}^{q} \beta_k X_{ik}.$$

Let $\boldsymbol{\beta}$ denote the q-dimensional vector with coordinates β_k, $1 \leq k \leq q$. The estimate a of α and the estimates b_k of β_k minimize the residual sum of squares

$$S_r = \sum_{i=1}^{n} r_i^2,$$

where

$$r_i = \bar{Y}_i - \hat{T}_i$$

and

$$\hat{T}_i = a + \sum_{k=1}^{q} b_k X_{ik}.$$

The estimates a and b_k are uniquely determined if the sample covariance matrix $\widehat{\mathrm{Cov}}(\mathbf{X})$ of the $\mathbf{X}_i$, $1 \leq i \leq n$, is positive definite. In case the estimates are not unique, then they may be chosen both to minimize S_r and to minimize $a^2 + \mathbf{b}'\mathbf{b}$ (Rao & Mitra, 1971, p. 51). Under the above mentioned assumptions on the rater errors e_{ij}, the mean-squared error of

$$d_{Ti} = T_i - T_i^*$$

is

$$\sigma_{dT}^2 = E([d_{Ti}]^2),$$

and

$$\sigma_d^2 = E(\bar{Y}_i - T_i + T_i - T_i^*)^2 = E(T_i - T_i^*)^2 + E(\bar{Y}_i - T_i)^2 = \sigma_{dT}^2 + \sigma^2/m_H, \tag{2}$$

where $m_H = 1/E(1/m_i)$ is the harmonic mean of m_i. If each m_i is m, then $E(1/m_i) = 1/m$ (Haberman, 2006). The mean-squared error σ_d^2 defined in Equation 1 is the smallest mean-squared error achievable by linear prediction of $\bar{Y}_i$ by $\mathbf{X}_i$ when the joint covariance matrix of $\bar{Y}_i$ and $\mathbf{X}_i$ is known. Similarly, σ^2_{dT} is the smallest mean-squared error achievable by linear prediction of the true essay score T_i by $\mathbf{X}_i$. Conditional on the use of M raters, so that $m_i = M > 0$, the smallest mean-squared error achievable by linear prediction of $\bar{Y}_i$ by $\mathbf{X}_i$ is

$$\sigma^2_{dM} = E(d_i^2|m_i = M) = \sigma^2_{dT} + \sigma^2/M.$$

To judge the effectiveness of linear regression, it is often helpful to compare the mean-squared error achieved by a trivial prediction of $\bar{Y}_i$ or T_i in which a constant predictor α is used. The best choice of α for both cases is $E(\bar{Y}_i) = E(T_i)$. The mean-squared error for prediction of $\bar{Y}_i$ by $E(\bar{Y}_i)$ is the variance σ^2_{Yb} of $\bar{Y}_i$, and the mean-squared error for prediction of T_i by $E(T_i)$ is the variance σ^2_T of T_i. Clearly

$$\sigma^2_{Yb} = \sigma^2_T + \sigma^2/m_H. \tag{3}$$

The proportional reduction of mean-squared error achieved by linear prediction of $\bar{Y}_i$ by T_i^* instead of by $E(\bar{Y}_i)$ is then

$$\rho^2_{Yb} = \frac{\sigma^2_{Yb} - \sigma^2_d}{\sigma^2_{Yb}}. \tag{4}$$

In the case of linear prediction of T_i, the proportional reduction of mean-squared error in predicting T_i by T_i^* instead of by $E(T_i)$ is

$$\rho^2_T = \frac{\sigma^2_T - \sigma^2_{d_T}}{\sigma^2_T} = \frac{\sigma^2_{Yb} - \sigma^2/m_H - \sigma^2_d + \sigma^2/m_H}{\sigma^2_T} = \frac{\sigma^2_{Yb} - \sigma^2_d}{\sigma^2_T} = \rho^2_{Yb}\sigma^2_{Yb}/\sigma^2_T,$$

using Equations 2, 3, and 4. Note that σ^2_T is less than σ^2_{Yb}, so that ρ^2_T exceeds ρ^2_{Yb}.

Conditional on $m_i = M > 0$, the mean-squared error for prediction of $\bar{Y}_i$ by $E(\bar{Y}_i)$ is

$$\sigma^2_{YbM} = E([\bar{Y}_i - E(\bar{Y}_i)]^2|m_i = M) = E([T_i - E(T_i)]^2) + E([\bar{Y}_i - T_i]^2)^2 = \sigma^2_T + \sigma^2/M,$$

so that the proportional reduction in mean-squared error in predicting $\bar{Y}_i$ by $\mathbf{X}_i$ instead of by $E(\bar{Y}_i)$ is

$$\rho^2_{YbM} = \frac{\sigma^2_{YbM} - \sigma^2_{dM}}{\sigma^2_{YbM}} = \rho^2_{Yb}\sigma^2_{Yb}/\sigma^2_{YbM}.$$

The relationship of ρ^2_{Yb} to ρ^2_{YbM} depends on whether M exceeds m_H.

3.2 Inflation of Mean-Squared Error

Cross-validation entails evaluation of the conditional mean-squared error

$$\tau_{r0}^2 = E(r_0^2|\mathbf{X}_i, 1 \le i \le n)$$

for prediction of $\bar{Y}_0$ by $\hat{T}_0$ given the predictors $\mathbf{X}_i$, $1 \le i \le n$. The important issue is that $r_0 = \bar{Y}_0 - \hat{T}_0$ is the prediction error of the average score $\bar{Y}_0$ based on the predicted average $\hat{T}_0$, where $\hat{T}_0$ employs the predictors from essay 0 but has estimates a and b_k developed from essays 1 to n. Let $f_i = \hat{T}_i - T_i^*$ be the difference between the estimated best linear predictor $\hat{T}_i$ and the actual best linear predictor T_i^*, and let

$$\tau_{f0}^2 = E(f_0^2|\mathbf{X}_i, 1 \le i \le n).$$

It can be shown that

$$\tau_{r0}^2 = \sigma_d^2 + \tau_{f0}^2 > \sigma_d^2. \tag{5}$$

As the sample size n becomes large, standard large-sample arguments as in Box (1954) and Gilula and Haberman (1994) can be used to prove that $n\tau_{f0}^2$ converges with probability 1 to

$$p = \sigma_d^2 + \mathrm{tr}([\mathrm{Cov}(\mathbf{X})]^{-1}\,\mathrm{Cov}(d\mathbf{X})), \tag{6}$$

where tr is the trace operator and $\mathrm{Cov}(d\mathbf{X})$ is the covariance matrix of $d_i\mathbf{X}_i$. In the special case in which d_{Ti} is independent of $\mathbf{X}_i$, $p = (q+1)\sigma_d^2$ (Haberman, 2006), and $n\tau_{f0}^2$ is p whenever the sample covariance matrix of the $\mathbf{X}_i$, $1 \le i \le n$, is positive definite. More generally, if c_1 is the minimum possible conditional expectation $E(d_i^2|\mathbf{X}_i)$ of d_i^2 given $\mathbf{X}_i$ and c_2 is the maximum possible conditional expectation $E(d_i^2|\mathbf{X}_i)$ of d_i^2 given $\mathbf{X}_i$, then p is between $(q+1)c_1$ and $(q+1)c_2$. Note that for d_i independent of $\mathbf{X}_i$, $c_1 = c_2 = \sigma_d^2$, so that the general result indeed implies that $p = (q+1)\sigma_d^2$.

Consider the relative increase

$$I = \tau_{r0}^2/\sigma_d^2 - 1$$

in conditional mean-squared error due to parameter estimation (i.e., due to estimation of $\bar{Y}_0$ by $\hat{T}_0$ instead of by T_0^*). This relative increase, which may be termed *inflation in* *mean-squared error*, plays a key role in the methodology considered in this report. The relative increase I is $(q+1)/n$ if, for each essay i, d_{Ti} is independent of $\mathbf{X}_i$ and $\mathbf{X}_i$ has a positive-definite sample covariance matrix. More generally, with probability 1, nI has a limit between $(q+1)c_1/c_2$ and $(q+1)c_2/c_1$. Note that $c_1 = c_2 = \sigma_d^2$ if the standard regression assumptions hold, so that the general formula is consistent with the fact that $nI = q+1$ if the sample covariance matrix of the $\mathbf{X}_i$, $1 \le i \le n$, is positive-definite. For instance, if $q = 8$ as in the e-rater example, then a sample size of 360 would be expected to yield a relative increase in mean-squared error of 2.5% if standard regression assumptions hold and the $\mathbf{X}_i$, $1 \le i \le n$, have a positive-definite covariance matrix.

Computations of relative increases in mean-squared error must be modified to some extent to study the increase prediction error for the true essay score T_i. In this case, consider the error

$$r_{Ti} = T_i - \hat{T}_i.$$

The conditional mean-squared error

$$\tau_{rT0}^2 = E([r_{T0}]^2 | \mathbf{X}_i, 1 \le i \le n)$$

is compared to $E([T_i - T_i^*]^2)^2 = \sigma_{dT}^2$. Because

$$\tau_{rT0}^2 = \sigma_{dT}^2 + \tau_{f0}^2,$$

the relative increase

$$I_T = \tau_{rT0}^2 / \sigma_{dT}^2 - 1$$

is equal to $I\sigma_d^2/\sigma_{dT}^2 > I$.

Similar arguments can be applied to the relative increase I_M in mean-squared error for approximation of $\bar{Y}_0$ conditional on $m_0 = M > 0$. Given $m_0 = M$, the conditional mean-squared error

$$\tau_{rM0}^2 = E([\bar{Y}_0 - \hat{T}_0]^2 | \mathbf{X}_i, 1 \le i \le n, m_0 = M) = \tau_{rT0}^2 + \sigma^2/M$$

for predicting $\bar{Y}_0$ by $\hat{T}_0$ is compared to the conditional mean-squared error σ_{dM}^2 for prediction of $\bar{Y}_0$ by T_0^*. The relative increase in mean-squared error, I_M, is defined as

$$I_M = \tau_{rM0}^2 / \sigma_{dM}^2 - 1.$$

3.3 Estimation of τ_{r0}^2 Using the Predicted Residual Sum of Squares (PRESS) Statistic

Estimation of mean-squared errors may be accomplished by use of deleted residuals or by some modification of standard results from the decomposition of sums of squares associated with regression analysis. The former approach is simpler to employ in terms of exploitation of commonly available software, although the latter approach is more efficient computationally. For each essay i, let $I(i)$ be the set of integers 1 to n that are not i. The *deleted residual* $d_{(i)}$ (Neter, Kutner, Nachtsheim, & Wasserman, 1996, pp. 372–373) is the difference $\bar{Y}_i - \hat{T}_{(i)}$, where

$$\hat{T}_{(i)} = a_{(i)} + \sum_{k=1}^{q} b_{k(i)} X_{ik}$$

and $a_{(i)}$ and $b_{k(i)}$ are found by minimizing the sum of squares

$$\sum_{j \in I(i)} \left[\bar{Y}_j - a_{(i)} - \sum_{k=1}^{q} b_{k(i)} X_{jk} \right]^2$$

in which data from essay i are omitted. Computation of $d_{(i)}$ involves minimal work, for

$$d_{(i)} = r_i/(1 - h_{ii}),$$

where

$$h_{ii} = n^{-1} + (\mathbf{X}_i - \bar{\mathbf{X}})'\mathbf{C}^{-1}(\mathbf{X}_i - \bar{\mathbf{X}})$$

is the ith diagonal element of the hat matrix (Draper & Smith, 1998, pp. 205–207), the vector of sample means of essay variables is

$$\bar{\mathbf{X}} = n^{-1} \sum_{i=1}^{n} \mathbf{X}_i,$$

and the matrix of corrected sums of cross products is

$$\mathbf{C} = \sum_{i=1}^{n} (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'.$$

Deleted residuals are commonly computed by standard software packages such as SAS. Given deleted residuals, τ_{r0}^2 may be estimated by the PRESS (Neter et al., 1996, pp.

345–346) sum of squares

$$s_{r0}^2 = n^{-1}\sum_{i=1}^{n} d_{(i)}^2.$$

The scaled difference $n[E(s_{r0}^2|\mathbf{X}_i, 1 \le i \le n) - \tau_{r0}^2]$ converges to 0 with probability 1, and $n^{1/2}(s_{r0}^2 - \tau_{r0}^2)$ converges in distribution to the variance of d_i^2. Alternatively, application of an expansion of $1/(1-h_{ii})^2$ shows that τ_{r0}^2 can be estimated by

$$\hat{\tau}_{r0}^2 = s_r^2\left(1 + \frac{2}{n}\right) + 2\,\mathrm{tr}([\widehat{\mathrm{Cov}}(\mathbf{X})]^{-1}\widehat{\mathrm{Cov}}(d\mathbf{X})), \tag{7}$$

where

$$s_r^2 = n^{-1}\sum_{i=1}^{n} r_i^2$$

is the residual sum of squares divided by n, $\widehat{\mathrm{Cov}}(\mathbf{X})$ is the sample covariance matrix of the $\mathbf{X}_i$, $1 \le i \le n$, and $\widehat{\mathrm{Cov}}(d\mathbf{X})$ is the sample covariance matrix of the $d_i\mathbf{X}_i$, $1 \le i \le n$. In case $\widehat{\mathrm{Cov}}(\mathbf{X})$ is singular, the Moore-Penrose inverse can be used. The scaled difference $n(s_{r0}^2 - \hat{\tau}_{r0}^2)$ converges in probability to 0.

3.4 Estimation of τ_{rT0}^2 *and* τ_{rM0}^2

Estimation of τ_{rT0}^2 may be accomplished if the probability is positive that some m_i exceeds 1. If m_i is not constant and if the conditional variance of e_{ij} given T_i or $\mathbf{X}_i$ is not assumed constant, then estimation is more complicated. A consistent estimate of σ^2 is provided by

$$\hat{\sigma}^2 = n_J^{-1}\sum_{i\in J}(m_i - 1)^{-1}\sum_{j=1}^{m_i}(Y_{ij} - \bar{Y}_i)^2,$$

where J is the set of integers i, $1 \le i \le n$, with $m_i > 1$, and n_J is the number of integers i in J. If all m_i are equal, then $\hat{\sigma}^2$ is just a within-groups mean-squared error from a one-way analysis of variance. If J is empty, then $\hat{\sigma}^2$ may be set to 0; however, such an estimate is obviously not satisfactory. As the sample size n becomes large, $\hat{\sigma}^2$ converges with probability 1 to σ^2. Given $\hat{\sigma}^2$, τ_{rT0}^2, which can be shown to be equal to $\tau_{r0}^2 - \sigma^2/m_H$, can be estimated by

$$s_{rT0}^2 = s_{r0}^2 - \hat{\sigma}_e^2,$$

where

$$\hat{\sigma}_e^2 = \hat{\sigma}^2/\hat{m}_H$$

and $\hat{m}_H$ is the sample harmonic mean of the m_i, $1 \le i \le n$. One may also substitute $\hat{\tau}_{r0}^2$ for s_{r0}^2. In like manner,

$$s_{rM0}^2 = s_{rT0}^2 + \hat{\sigma}^2/M$$

may be used to estimate τ_{rM0}^2, the conditional mean of r_0^2 given $m_0 = M$.

3.5 Estimation of σ_d^2, σ_{dT}^2, and Inflations of Mean-Squared Error

As proved in the appendix, an estimate of σ_d^2 is

$$s_d^2 = (s_r^2 + s_{r0}^2)/2.$$

The more conventional estimate $ns_r^2/(n-q-1)$ of σ_d^2 is not appropriate if the residual d_i and predictor $\mathbf{X}_i$ are not independent. In general $n^{1/2}(s_d^2 - \sigma_d^2)$ has the same normal approximation as $n^{1/2}(s_{r0}^2 - \tau_{r0}^2)$, and $n[E(s_d^2|\mathbf{X}_i, 1 \le i \le n) - \sigma_d^2]$ converges to 0 with probability 1. It then follows that the relative increase I in mean-squared error can be estimated by

$$\hat{I} = \frac{s_{r0}^2 - s_r^2}{s_{r0}^2 + s_r^2}.$$

It also follows that an estimate of the value of I achieved if n^* observations are used rather than n is $\hat{I}^* = n\hat{I}/n^*$. This estimate provides a guide to sample-size selection. Note that if the standard regression assumptions hold, then it can be asserted without performing any estimation that nI will be close to $q+1$. However, in a real application, where one rarely knows whether the regression assumptions are true, it is recommended that I be estimated by $\hat{I}$ and that the estimates $\hat{I}$ and $\hat{I}^*$ guide the process of sample-size selection.

If $m_i > 1$ with positive probability, then, using Equation 2, σ_{dT}^2 may be estimated by

$$s_{dT}^2 = s_d^2 - \hat{\sigma}_e^2,$$

and I_T may be estimated by

$$\hat{I}_T = \frac{s_{rT0}^2 - s_{dT}^2}{s_{dT}^2}.$$

Similar arguments apply to estimation of I_M. The estimate of σ^2_{dM} is $s^2_{dM} = s^2_{dT} + \hat{\sigma}^2/M$, so that I_M has estimate

$$\hat{I}_M = \frac{s^2_{rM0} - s^2_{dM}}{s^2_{dM}}.$$

3.6 Estimation of Proportional Reduction in Mean-Squared Error Using Cross-Validation

Proportional reduction in mean-squared error may also be considered in terms of cross-validation, for which results are especially simple when $\bar{Y}_0$ is approximated by the sample mean $\bar{Y}$ of the $\bar{Y}_i$, $1 \le i \le n$. The error $r_{0C} = \bar{Y}_0 - \bar{Y}$ has mean 0 and variance

$$\tau^2_{r0C} = \sigma^2_{Yb}(1 + n^{-1}).$$

Estimation can be accomplished by use of the conventional estimate

$$s^2_{Yb} = (n-1)^{-1} \sum_{i=1}^{n} (\bar{Y}_i - \bar{Y})^2$$

for σ^2_{Yb}, so that τ^2_{r0C} is estimated by

$$\hat{\tau}^2_{r0C} = s^2_{r0C}(1 + n^{-1}).$$

Use of deleted residuals results in the estimate

$$s^2_{r0C} = [n^2/(n^2-1)]\hat{\tau}^2_{r0C}.$$

Thus for estimation of $\bar{Y}_0$, the proportional reduction

$$\rho^2_{Yb0} = \frac{\tau^2_{r0C} - \tau^2_{r0}}{\tau^2_{r0C}}$$

in mean-squared error achieved by linear prediction of $\bar{Y}_0$ by $\hat{T}_0$ instead of by $\bar{Y}$ is estimated by

$$\hat{\rho}^2_{Yb0} = \frac{s^2_{r0C} - s^2_{r0}}{s^2_{r0C}}.$$

As the sample size n increases, $\hat{\rho}^2_{Yb0}$ converges with probability 1 to ρ^2_{Yb}, and ρ^2_{Yb0} converges to ρ^2_{Yb}. Similar approximations are available for ρ^2_{T0} and ρ^2_{Yb0M}. The mean square of

$r_{T0C} = T_0 - \bar{Y}$ is $\tau^2_{rT0C} = \tau^2_{r0C} - \sigma^2/m_H$, and the conditional variance of r_{0C} given $m_i = M$ is $\tau^2_{rM0C} = \tau^2_{rT0C} + \sigma^2/M$. The estimate of τ^2_{rT0C} is then $s^2_{rT0C} = s^2_{r0C} - \hat{\sigma}^2/\hat{m}_H$, and the estimate of τ^2_{rM0C} is $s^2_{rT0C} + \hat{\sigma}^2/M$. Hence,

$$\rho^2_{T0} = \frac{\tau^2_{rT0C} - \tau^2_{rT0}}{\tau^2_{rT0C}},$$

the proportional reduction in mean-squared error achieved by linear prediction of T_0 by $\hat{T}_0$ instead of by $\bar{Y}$, is estimated by

$$\hat{\rho}^2_{T0} = \hat{\rho}^2_{Yb0}\frac{s^2_{r0C}}{s^2_{rT0C}},$$

and

$$\rho^2_{M0} = \frac{\tau^2_{rM0C} - \tau^2_{rM0}}{\tau^2_{rM0C}}$$

is estimated by

$$\hat{\rho}^2_{M0} = \hat{\rho}^2_{Yb0}\frac{s^2_{r0C}}{s^2_{rM0C}}.$$

3.7 A Practitioner's Guide: What Quantities To Examine in a Real Application?

Table 3 lists all the residuals, sums of squares, inflations of mean-squared error, and proportional reductions described above. An important question, given so many different quantities, is the following: Which of these quantities should we use and how in a real application? The answers are quite closely linked to answers found in regression analysis, and depends on the goal of the user of the methodology. We will discuss three potential users and describe the quantities each would be interested in. Estimates and interpretations are considered for the first prompt in the second group to illustrate their application.

User 1: One who wants an idea of the errors when two raters are used. The parameter σ^2_d measures the mean-squared error for prediction of an average holistic score by observed essay features in the case in which the regression coefficients are known. Thus σ^2_d is a lower bound on the mean-squared error that can possibly be achieved when the regression coefficients are estimated from a sample of essays. If σ^2_d is regarded as too large for the application, then no amount of sampling can lead to a satisfactory linear prediction of the average holistic score. For the above mentioned prompt, the estimate s^2_d of σ^2_d may be found

Table 3.

All Residuals, Sums of Squares, Variance Inflations, and Proportional Reductions

Quantity	Definition	Notes
d_i	$\bar{Y}_i - T_i^*$	
r_i	$\bar{Y}_i - \hat{T}_i$	
d_{Ti}	$T_i - T_i^*$	
r_{Ti}	$T_i - \hat{T}_i$	
r_{0C}	$\bar{Y}_0 - \bar{Y}$	
r_{T0C}	$T_0 - \bar{Y}$	
σ_d^2	$E(d_i^2)$	$\sigma_d^2 = \sigma_{dT}^2 + \sigma^2/m_H$
σ_{dT}^2	$E(d_{Ti}^2)$	
σ_{dM}^2	$E(d_i^2 \vert m_i = M)$	$\sigma_{dM}^2 = \sigma_{dT}^2 + \sigma^2/M$
σ_{Yb}^2	$E([\bar{Y}_i - E\bar{Y}_i]^2)$	$\sigma_{Yb}^2 = \sigma_T^2 + \sigma^2/m_H$
σ_{YbM}^2	$E([\bar{Y}_i - E\bar{Y}_i]^2 \vert m_i = M)$	$\sigma_{YbM}^2 = \sigma_T^2 + \sigma^2/M$
τ_{r0}^2	$E(r_0^2 \vert \mathbf{X}_i, 1 \leq i \leq n)$	
τ_{rT0}^2	$E(r_{T0}^2 \vert \mathbf{X}_i, 1 \leq i \leq n)$	
τ_{rM0}^2	$E(r_0^2 \vert m_0 = M, \mathbf{X}_i, 1 \leq i \leq n)$	$\tau_{rM0}^2 = \tau_{rT0}^2 + \sigma^2/M$
τ_{r0C}^2	$E(r_{0C}^2)$	$\tau_{r0C}^2 = \sigma_{Yb}^2(1 + 1/n)$
τ_{rT0C}^2	$E(r_{T0C}^2)$	$\tau_{rT0C}^2 = \tau_{r0C}^2 - \sigma^2/m_H$
τ_{rM0C}^2	$E(r_{0C}^2 \vert m_0 = M, \mathbf{X}_i, 1 \leq i \leq n)$	$\tau_{rM0C}^2 = \tau_{rT0C}^2 + \sigma^2/M$
I	$\tau_{r0}^2/\sigma_d^2 - 1$	RI:$\bar{Y}_0$, $\hat{T}_0$, T_0^*
I_T	$\tau_{rT0}^2/\sigma_{dT}^2 - 1$	RI: T_0, $\hat{T}_0$, T_0^*
I_M	$\tau_{rM0}^2/\sigma_{dM}^2 - 1$	Given $m_0 = M$, RI: $\bar{Y}_0$, $\hat{T}_0$, T_0^*
ρ_{Yb}^2	$(\sigma_{Yb}^2 - \sigma_d^2)/\sigma_{Yb}^2$	PRMSE: $\bar{Y}_i$, T_i^*, $E(\bar{Y}_i)$
ρ_T^2	$(\sigma_T^2 - \sigma_{dT}^2)/\sigma_T^2$	PRMSE: T_i, T_i^*, $E(T_i)$
ρ_{YbM}^2	$\rho_{YbM}^2 = \rho_{Yb}^2\sigma_{Yb}^2/\sigma_{YbM}^2$	Given $m_0 = M$, PRMSE: $\bar{Y}_i$, T_i^*, $E(\bar{Y}_i)$
ρ_{Yb0}^2	$(\tau_{r0C}^2 - \tau_{r0}^2)/\tau_{r0C}^2$	PRMSE: $\bar{Y}_0$, $\hat{T}_0$, $\bar{Y}$
ρ_{T0}^2	$(\tau_{rT0C}^2 - \tau_{rT0}^2)/\tau_{rT0C}^2$	PRMSE: T_0, $\hat{T}_0$, $\bar{Y}$
ρ_{M0}^2	$(\tau_{rM0C}^2 - \tau_{rM0}^2)/\tau_{rM0C}^2$	Given $m_0 = M$, PRMSE: $\bar{Y}_0$, $\hat{T}_0$, $\bar{Y}$

Note. "PRMSE: a, b, c" means the proportional reduction in mean-squared error by prediction of a by b compared to prediction of a by c. "RI: d, e, f" means the relative increase in mean-squared error, conditional on $\mathbf{X}_i$, $1 \leq i \leq n$, due to estimation of d by e compared to estimation of d by f.

in the following fashion from regression output from SAS. The estimated root mean-squared error

$$\left[(n-q-1)^{-1}\sum_{i=1}^{n} r_i^2\right]^{1/2} = 0.5725,$$

where n is 373 and q is 8. It follows that

$$s_r^2 = n^{-1}\sum_{i=1}^{n} r_i^2 = \frac{n-q-1}{n}(0.5725)^2 = 0.3198.$$

The PRESS sum of squares

$$\sum_{i=1}^{2} d_{(i)}^2 = 125.5,$$

so that

$$s_{r0}^2 = n^{-1}\sum_{i=1}^{n} d_{(i)}^2 = 0.3364.$$

It follows that

$$s_d^2 = (s_r^2 + s_{r0}^2)/2 = (0.3198 + 0.3364)/2 = 0.3281.$$

By itself, this estimate does not suggest a precise approximation to the average human score, for the square root s_d of s_d^2 is 0.5728, and the underlying measurements are on a 6-point scale. Nonetheless, alternatives must be considered to provide a proper perspective.

The simplest alternative measure is σ_{Yb}^2, the mean-squared error for prediction of the average holistic score when the mean holistic score is known and no essay features are employed in the prediction. The estimate s_{Yb}^2 is the square 0.9496 of the sample standard deviation 0.9745 of the $\bar{Y}_i$ that is reported by SAS. Thus the estimate s_d^2 of the mean-squared error σ_d^2 is somewhat smaller than is the estimate s_{Yb}^2 associated with a trivial constant predictor. The coefficient ρ_{Yb}^2 then is the proportional reduction in mean-squared error achieved by prediction of the average holistic score by use of a regression on essay features in which all population means, variances, and covariances are known. The estimate of ρ_{Yb}^2 is

$$1 - s_d^2/s_{Yb}^2 = 1 - 0.3281/0.9496 = 0.6545,$$

so that the regression analysis is predicted to be much more effective than use of a constant predictor.

In practice, using a sample of essays to estimate population parameters leads to less satisfactory results than are obtained if population parameters are known. Thus τ^2_{r0} is the mean-squared error, conditional on the observed essay features in the sample, achieved when predicting an average holistic score for an essay not in the sample by use of essay features with regression parameters estimated by the observed sampling data. If data suggest that τ^2_{r0} is excessive for the application but σ^2_d is acceptable, then using a larger sample is appropriate. The absolute loss in mean-squared error due to sampling is $\tau^2_{r0} - \sigma^2_d$. If the estimated value of σ^2_d is acceptably small, then the sample size required for τ^2_{r0} to have an estimate that is acceptably small can be estimated. In the example, τ^2_{r0} is estimated by $s^2_{r0} = 0.3364$. As expected, s^2_{r0} is larger than the estimated mean-squared error $s^2_d = 0.3281$. The loss of precision due to estimation of regression coefficients is quantified in the coefficient I, the relative loss in mean-squared error due to sampling. The estimated value of I is

$$\hat{I} = (s^2_{r0} - s^2_d)/(s^2_d) = (0.3364 - 0.3281)/0.3281 = 0.02518,$$

so that the relative increase in mean-squared error is about 2.5%. This estimated relative increase is not surprising, for $(q+1)/n = (8+1)/373 = 0.02413$. Such an increase might well be considered acceptable. The absolute increase $0.3364 - 0.3281 = 0.008262$ also appears acceptably small.

Given that a sample of essays is used to estimate prediction parameters, an added measure of interest is ρ^2_{Yb0}, the proportional reduction in mean-squared error achieved by prediction of average holistic scores by essay features in a sample of essays. The coefficient ρ^2_{Yb0} is normally less than ρ^2_{Yb}, and the difference between the two coefficients provides an added measure of the loss of predictive power due to the effects of parameter estimation from a sample. For the example, ρ^2_{Yb0} is estimated by

$$\hat{\rho}^2_{Yb0} = 1 - s^2_{r0}/s^2_{r0C},$$

where

$$s^2_{r0C} = s^2_{Yb}\frac{n}{n-1} = 0.9496(373/372) = 0.9522.$$

Thus

$$\hat{\rho}^2_{Yb0} = 1 - 0.3364/0.9522 = 0.6467.$$

and the proportional reduction of mean-squared error is about two-thirds. Comparison to the estimate of ρ^2_{Yb} suggests that the loss due to the estimation of regression coefficients and means is relatively small. *User 2: One who wants to know what proportion of the error in estimation can be attributed to the raters.* This user will be interested in the results that can be achieved if rater error disappears. This analysis permits the practitioner to distinguish between prediction error that is inevitable given rater error and prediction error that results from an imperfect relationship between the true essay score and the essay features. Thus σ^2_{dT} measures the prediction error of the true essay score by the essay features when all needed means, variances, and covariances are known, and τ^2_{rT0} is the corresponding measure when regression parameters are estimated from the sample and the essay under study is not in the sample. For estimation of these measures, one uses the estimated rater variance

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \sum_{j=1}^{2} (Y_{ij} - \bar{Y}_i)^2.$$

Here formulas simplify because each essay has two raters, so that each $m_i = 2$, J is the set of integers from 1 to $n = 373$, and $n_J = n$. The estimate $\hat{\sigma}^2 = 0.1260$, so that $\hat{\sigma}^2_e$, the estimated variance of the average rater error $\bar{e}_i$ for essay i, is $0.1260/2 = 0.0630$. It follows that the mean-squared error σ^2_{dT} has the estimate

$$s^2_{dT} = s^2_d - \hat{\sigma}^2_e = 0.3281 - 0.0630 = 0.2651.$$

Thus a substantial fraction (0.0630/0.3281=0.192) of the estimated mean-squared error s^2_d for prediction of the average holistic score $\bar{Y}_i$ is due to rater variability. In like fashion, τ^2_{rT0} has the estimate

$$s^2_{rT0} = s^2_{r0} - \hat{\sigma}^2_e = 0.3364 - 0.0630 = 0.2734.$$

No matter how large the sample may be, τ^2_{rT0} cannot be less than σ^2_{dT}. The difference between τ^2_{rT0} and σ^2_{dT} is the same as the difference between τ^2_{r0} and σ^2_d; however, σ^2_{dT} is less than σ_d^2, so that the proportional increase I_T in mean-squared error for predicting the true essay score if sampling is used to estimate regression parameters is greater than the corresponding proportional increase I for prediction of the average holistic score. In the example, I_T has estimate

$$\hat{I}_T = (0.2734 - 0.2651)/0.2651 = 0.03117,$$

so that the inflation of mean-squared error of about 3% is somewhat larger for predicting true holistic scores than was the case for predicting average holistic scores.

The proportional reduction ρ_T^2 in mean-squared error for predicting true essay score by essay features when all means, covariances, and variances are known is larger than the corresponding proportional reduction ρ_{Yb}^2 in mean-squared error for predicting average essay scores by essay features. When sampling is required, the proportional reduction ρ_{T0}^2 in mean-squared error for predicting true essay scores by essay features is normally smaller than ρ_T^2. A striking aspect of automated essay scoring is that ρ_T^2 and ρ_{T0}^2 can be quite high, say 0.9. In the example, results are less striking. With use of a constant predictor, the estimated mean-squared error for prediction of T_0 is

$$s_{rT0C}^2 = s_{r0C}^2 - \hat{\sigma}_e^2 = 0.9522 - 0.0630 = 0.8892,$$

so that the estimated proportional reduction in mean-squared error is

$$\hat{\rho}_{T0}^2 = (0.8892 - 0.2734)/0.8892 = 0.6926.$$

Thus the regression analysis has accounted for about 70% of the mean-squared error that is not due to rater variability. *User 3: One who wants an idea of the errors when one rater is employed from the data based on two raters per essay.* In some cases, a testing program may consider using an automated score in place of some fixed number M of human ratings of an essay. Coefficients σ_{dM}^2, τ_{rM0}^2, I_M, ρ_{YbM}^2, and ρ_{M0}^2 are provided for this case. They are interpreted as in the case of ordinary prediction of the average holistic score given the added condition that the number of raters is specified to be M. Illustrations in this report use $M = 1$. For the example, the estimated mean-squared error s_{dM}^2 for known population characteristics is

$$s_{dT}^2 + \hat{\sigma}^2 = 0.2651 + 0.1260 = 0.3911,$$

while the estimated cross-validation mean-squared error derived from deleted residuals is

$$s_{rM0}^2 = s_{rT0}^2 + \hat{\sigma}^2 = 0.3994.$$

Note that, due to use of only one rater rather than two, the predictions of average holistic scores are somewhat less accurate here than in the original case of two raters. The corresponding estimated cross-validation mean-squared error with a constant predictor is

$$s_{rM0C}^2 = s_{rT0C}^2 + \hat{\sigma}^2 = 0.8892 + 0.1260 = 1.015,$$

so that the estimated proportional reduction in mean-squared error is

$$\hat{\rho}_{M0}^2 = (1.015 - 0.3994)/1.015 = 0.6066.$$

The reduced proportional reduction in mean-squared error for one rather than two raters is predictable. It is also predictable that inflation of mean-squared error is reduced compared to $\hat{I}$ and $\hat{I}_T$ for this case. The inflation I_M is estimated by

$$\hat{I}_M = (0.3994 - 0.3911)/0.3994 = 0.02113.$$

Thus the inflation of mean-squared error for one rater is about 2% rather than the approximate 2.5% achieved for two raters.

3.8 Results From Analysis of the Data

To begin, each essay in the four groups of prompts was analyzed. A summary of results is reported in Table 4. Note that results for Group 1 are omitted for entries that rely on $\hat{\sigma}^2$ due to the very limited number of essays that have been double-scored. The case of $M = 1$ is considered in the table. Several basic conclusions appear possible, at least for these groups of prompts. In typical cases, the means of the estimates $n\hat{I}$ of inflation of mean-squared error are comparable to the ideal value of 9 associated with an intercept and 8 predictors (i.e., the inflation of mean-squared error due to estimation is roughly

Table 4.

Summary of Regression Analysis of Individual Essay Prompts Within Groups

	Group 1		Group 2		Group 3		Group 4	
Statistic	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
n	495.9	6.5	456.0	62.0	467.0	14.4	499.8	6.2
s^2_{r0}	0.352	0.061	0.367	0.114	0.289	0.053	0.496	0.201
s^2_{r0C}	0.871	0.084	1.234	0.255	1.688	0.322	1.215	0.297
$\hat{I}$	0.0198	0.0019	0.0225	0.0049	0.0234	0.0028	0.0203	0.0022
$n\hat{I}$	9.81	0.91	10.05	1.29	10.94	1.18	10.18	1.14
$\hat{\rho}^2_{Yb0}$	0.593	0.075	0.687	0.131	0.826	0.027	0.602	0.092
s^2_{rT0}			0.309	0.121	0.173	0.052	0.361	0.204
s^2_{rT0C}			1.175	0.261	1.572	0.323	1.080	0.301
$\hat{I}_T$			0.0286	0.0110	0.0415	0.0070	0.0319	0.0081
$n\hat{I}_T$			12.69	3.88	19.42	3.69	15.96	4.06
$\hat{\rho}^2_{T0}$			0.723	0.137	0.888	0.021	0.686	0.119
s^2_{r10}			0.426	0.112	0.404	0.058	0.632	0.198
s^2_{r10C}			1.292	0.251	1.803	0.322	1.351	0.293
$\hat{I}_1$			0.0190	0.0033	0.0166	0.0026	0.0154	0.0023
$n\hat{I}_1$			12.69	3.88	7.74	1.10	7.72	1.16
$\hat{\rho}^2_{10}$			0.655	0.128	0.771	0.039	0.537	0.078

similar to the amount anticipated if the standard assumptions of regression are valid). Typical estimated inflation $\hat{I}$ of mean-squared error is about 2%, a modest value. Even if typical sample sizes were halved to around 250, the inflation would be doubled to around 4%, a value that could be regarded as tolerable. The available estimates of proportional reductions in mean-squared error, $\hat{\rho}^2_{Yb0}$, $\hat{\rho}^2_{T0}$, and $\hat{\rho}^2_{10}$, indicate that families of prompts vary quite substantially as to how well e-rater predicts human scores, with the best results for the second and third groups of prompts.

3.9 Combining Essays in Groups

An alternative approach summarizes the data by looking at the prediction of a score for a group of essays. The initial approach is to use distinct regression coefficients for each prompt, so that K prompts in effect have $K(q+1)$ predictors. As evident from Table 5, results for this approach are quite similar to those for Table 4, except that the proportional

Table 5.
Summary of Regression Analysis of Essay Prompts Within Groups: Distinct Coefficients for Each Essay

Statistic	Group 1	Group 2	Group 3	Group 4
n	36,697	6,384	12,143	7,997
s^2_{r0}	0.352	0.366	0.288	0.496
s^2_{r0C}	0.892	1.320	1.698	1.355
$\hat{I}$	0.0198	0.0217	0.0235	0.0202
$n\hat{I}$	727.62	138.68	285.80	161.45
$\hat{\rho}^2_{Yb0}$	0.605	0.723	0.830	0.634
s^2_{rT0}	0.128	0.307	0.172	0.360
s^2_{rT0C}	0.668	1.262	1.582	1.219
$\hat{I}_T$	0.0568	0.0259	0.0399	0.0280
$n\hat{I}_T$	2,083.56	165.54	484.46	224.04
$\hat{\rho}^2_{T0}$	0.809	0.756	0.891	0.705
s^2_{r10}	0.366	0.424	0.403	0.632
s^2_{r10C}	0.906	1.378	1.813	1.491
$\hat{I}_1$	0.0191	0.0187	0.0167	0.0158
$n\hat{I}_1$	700.56	119.35	202.68	126.20
$\hat{\rho}^2_{10}$	0.596	0.693	0.777	0.576

reduction in mean-squared error is increased slightly. This is because it is computed relative to a constant predictor for all essays in the entire family rather than relative to a constant predictor for each prompt. The number of prompts scored in Group 1 is sufficient for analysis related to true scores; however, results for true scores and for exactly one score should be approached with caution given that the sampling assumptions appear questionable. The increase in the estimated product of sample size by relative inflation of mean-squared error primarily reflects the increased number of predictors present in the analysis.

Another approach of interest involves a much smaller number of predictors. A linear model is used for each group in which a separate intercept is used for each prompt, but the regression coefficients for the predictors are the same for each prompt. Results are summarized in Table 6. Relative to use of distinct intercepts and regression slopes for each prompt, estimated losses in mean-squared error are very limited (Groups 1, 2, and 4) or

Table 6.
Summary of Regression Analysis of Essay Prompts Within Groups: Distinct Intercepts for Each Essay, Common Regression Slopes

Statistic	Group 1	Group 2	Group 3	Group 4
n	36,697	6,384	12,143	7,997
s^2_{r0}	0.360	0.380	0.285	0.503
s^2_{r0C}	0.892	1.320	1.698	1.355
$\hat{I}$	0.0023	0.0036	0.0030	0.0032
$n\hat{I}$	82.83	22.80	36.03	25.25
$\hat{\rho}^2_{Yb0}$	0.597	0.712	0.832	0.629
s^2_{rT0}	0.135	0.322	0.169	0.368
s^2_{rT0C}	0.668	1.262	1.582	1.219
$\hat{I}_T$	0.0060	0.0042	0.0050	0.0043
$n\hat{I}_T$	221.50	26.93	60.68	34.62
$\hat{\rho}^2_{T0}$	0.798	0.745	0.893	0.699
s^2_{r10}	0.373	0.4384	0.400	0.639
s^2_{r10C}	0.906	1.378	1.813	1.491
$\hat{I}_1$	0.0022	0.0031	0.0021	0.0025
$n\hat{I}_1$	79.86	19.77	25.63	19.87
$\hat{\rho}^2_{10}$	0.588	0.682	0.779	0.571

nonexistent (Group 3). This approach has much more modest sample-size requirements than does the approach with individual regression coefficients for each prompt. If the group contains a substantial number of prompts, then a tolerable inflation of mean-squared error is obtained with about a tenth of the essays required with individual regression coefficients for each prompt. If the standard regression model applies, then the inflation of mean-squared error is approximately the number of prompts plus the number of predictors divided by the group sample size. In typical applications, the inflation will be approximated by one over the number of essays per prompt.

An even simpler model for a group of essays ignores the prompt entirely, so that the same intercepts and the same regression coefficients are applied to each essay in the group. Results are summarized in Table 7. Although the inflations of mean-squared error are very small, there is a substantial increase in the actual mean-squared error in Group 1 and in Group 2. Losses in mean-squared error are also encountered in the other groups, but they

Table 7.
Summary of Regression Analysis of Essay Prompts Within Groups: Common Intercepts and Common Regression Slopes

Statistic	Group 1	Group 2	Group 3	Group 4
n	36,697	6,384	12,143	7,997
s^2_{r0}	0.398	0.441	0.303	0.558
s^2_{r0C}	0.892	1.320	1.698	1.355
$\hat{I}$	0.0003	0.0015	0.0009	0.0013
$n\hat{I}$	9.36	9.44	10.78	10.10
$\hat{\rho}^2_{Yb0}$	0.554	0.666	0.829	0.608
s^2_{rT0}	0.173	0.383	0.175	0.396
s^2_{rT0C}	0.668	1.262	1.582	1.219
$\hat{I}_T$	0.0006	0.0017	0.0015	0.0017
$n\hat{I}_T$	21.50	10.87	17.89	13.56
$\hat{\rho}^2_{T0}$	0.740	0.697	0.889	0.675
s^2_{r10}	0.412	0.499	0.406	0.668
s^2_{r10C}	0.906	1.378	1.813	1.491
$\hat{I}_1$	0.0002	0.0013	0.0006	0.0010
$n\hat{I}_1$	9.05	8.34	7.72	8.04
$\hat{\rho}^2_{10}$	0.546	0.638	0.776	0.552

are very small in Group 3 and modest in Group 4. The one virtue of the approach with a common regression equation for each prompt is that the sample-size requirements for the group are similar to those for a single prompt. Thus one could consider use of several hundred essays for a complete group.

4. Sample-Size Determination for a Cumulative Logit Model

An alternative to linear regression analysis is cumulative logit analysis (Bock, 1973; Feng, Dorans, Patsula, & Kaplan, 2003; Haberman, 2006; McCullagh & Nelder, 1989; Pratt, 1981). This alternative has the advantage that approximations of $\bar{Y}_i$ and T_i must be within the range of possible essay scores. A mild disadvantage is that cross-validation is more difficult to perform with commonly available software.

4.1 The Statistical Model

The cumlative logit model assumes that, for each essay i, conditional on $\mathbf{X}_i$ and m_i, the scores Y_{ij}, $1 \leq j \leq m_i$, are independent random variables. For unknown parameters η_g, $1 \leq g < G$, and γ_k, $1 \leq k \leq q$, the conditional probability P_{ig} that $Y_{ij} \leq g$, $1 \leq g < G$, given $\mathbf{X}_i$, satisfies the cumulative logit relationship

$$\lambda_{ig} = \log(P_{ig}/(1 - P_{ig})) = \eta_g + \sum_{k=1}^{q} \gamma_k X_{ik}.$$

The method of analysis in this section does not assume that the probability model is true, just as the regression analysis in the previous section did not assume that the standard assumptions of linear regression were valid. Let $P_{iG} = 1$ and $P_{i0} = 0$. The η_g and γ_k are unique parameters defined to minimize the expected value of the average logarithmic penalty $\bar{L}_i$, where $\bar{L}_i$ is the average of L_{ij}, $1 \leq j \leq m_i$, and

$$L_{ij} = -\log[P_{ig} - P_{i(g-1)}]$$

if $Y_{ij} = g$ (Haberman, 1989; Gilula & Haberman, 1994). The cumulative logit model is evaluated by considering the mean-squared error from approximation of $\bar{Y}_i$ by its corresponding approximated expected value

$$T^*_{iL} = \sum_{g=1}^{G} g[P_{ig} - P_{i(g-1)}] = 1 + \sum_{g=1}^{G-1}(1 - P_{ig})$$

given $\mathbf{X}_i$. Maximum likelihood may be applied to the $\mathbf{X}_i$ and Y_{ij}, $1 \leq j \leq m_i$, for $1 \leq i \leq n$, to obtain estimates $\hat{\eta}_g$ of η_g and $\hat{\gamma}_k$ of γ_k. If maximum-likelihood estimates exist and the sample covariance matrix of the $\mathbf{X}_i$, $1 \leq i \leq n$, is nonsingular, then they are uniquely defined. Common statistical packages such as SAS may be employed for this purpose. Given the parameter estimates, one may estimate λ_{ig} by

$$\hat{\lambda}_{ig} = \hat{\eta}_g + \sum_{k=1}^{q} \hat{\gamma}_k X_{ik}$$

and P_{ig} by

$$\hat{P}_{ig} = [1 + \exp(-\hat{\lambda}_{ig})]^{-1}.$$

The true score T_i and the mean $\bar{Y}_i$ are both estimated by

$$\hat{T}_{iL} = \sum_{g=1}^{G} g[\hat{P}_{ig} - \hat{P}_{i(g-1)}] = 1 + \sum_{g=1}^{G-1}(1 - \hat{P}_{ig}).$$

The η_g and γ_k are uniquely defined as long as the conditional probabilities P_{ig} are strictly increasing in g for fixed i and the covariance matrix of $\mathbf{X}_i$ is positive definite.

An analysis quite similar to that for linear prediction may be considered. The major change involves cross-validation. To be consistent with the error criterion used with least squares, consider the mean-squared error

$$\sigma^2_{dL} = E(d^2_{iL}),$$

where

$$d_{iL} = \bar{Y}_i - T^*_{iL}.$$

The residual sum of squares is

$$S_{rL} = \sum_{i=1}^{n} r^2_{iL}$$

for

$$r_{iL} = \bar{Y}_i - \hat{T}_{iL}.$$

The mean-squared error of $d_{TiL} = T_i - T^*_{iL}$ is $\sigma^2_{dTL} = E([d_{TiL}]^2)$, and

$$\sigma^2_{dL} = \sigma^2_{dTL} + \sigma^2/m_H. \tag{8}$$

Conditional on the use of M raters, so that $m_i = M > 0$, the conditional mean-squared error of T^*_{iL} as a predictor of $\bar{Y}_i$ is

$$\sigma^2_{dML} = E(d^2_{iL}|m_i = M) = \sigma^2_{dTL} + \sigma^2/M.$$

If the cumulative logit model is correct, then the proposed estimation approach is efficient. If the model is not correct, then it may be the case that η'_g and γ'_k can be found such that

$$E([\bar{Y}_i - T'_{iL}]^2) < \sigma^2_{dL}$$

for

$$T'_{iL} = \sum_{g=1}^{G} g[P'_{ig} - P'_{i(g-1)}] = 1 + \sum_{g=1}^{G-1}(1 - P'_{ig}),$$

$$P'_{ig} = [1 + \exp(-\lambda'_{ig})]^{-1},$$

and

$$\lambda'_{ig} = \eta'_g + \sum_{k=1}^{q} \gamma'_k X_{ik}.$$

Given a desire to employ common software in a routine fashion, no attempt has been made to exploit this possibility.

The proportional reduction of mean-squared error achieved by linear prediction of $\bar{Y}_i$ by T^*_{iL} instead of by $E(\bar{Y}_i)$ is

$$\rho^2_{YbL} = \frac{\sigma^2_{Yb} - \sigma^2_{dL}}{\sigma^2_{Yb}}. \tag{9}$$

The proportional reduction of mean-squared error in predicting T_i by T^*_{iL} instead of by $E(T_i)$ is

$$\rho^2_{TL} = \frac{\sigma^2_T - \sigma^2_{d_{TL}}}{\sigma^2_T} = \rho^2_{YbL}\sigma^2_{Yb}/\sigma^2_T.$$

Because σ^2_T is less than σ^2_{Yb}, ρ^2_{TL} exceeds ρ^2_{YbL}. Given that $m_i = M$, the proportional reduction in mean-squared error in predicting T_i by $\bar{Y}_i$ instead of by $E(T_i)$ is

$$\rho^2_{YbML} = \rho^2_{YbL}\sigma^2_{Yb}/\sigma^2_{YbM}.$$

The relationship of ρ^2_{YbL} to ρ^2_{YbML} depends on whether M exceeds m_H.

4.2 Inflation of Mean-Squared Error

In the case of cumulative logit analysis, cross-validation entails evaluation of the conditional mean-squared error $\tau^2_{r0L} = E(r^2_{0L}|\mathbf{X}_i, 1 \le i \le n)$ for prediction of $\bar{Y}_0$ by $\hat{T}_{0L}$ given the predictors $\mathbf{X}_i$, $1 \le i \le n$. Let $f_{iL} = \hat{T}_{iL} - T^*_{iL}$ be the difference between the estimated predictor $\hat{T}_{iL}$ and the actual predictor T^*_{iL}, and let τ^2_{f0L} be the conditional expected value $E(f^2_{0L}|\mathbf{X}_i, 1 \le i \le n)$ of the squared deviation f^2_{0L} given the predictors $\mathbf{X}_i$, $1 \le i \le n$. Then

$$\tau^2_{r0L} = \sigma^2_{dL} + \tau^2_{f0L} > \sigma^2_{dL}.$$

As the sample size n becomes large, standard large-sample arguments similar to those for log-linear models (Gilula & Haberman, 1994) show that $n\tau^2_{f0L}$ converges with probability 1 to a constant p_L. Of interest is the relative increase

$$I_L = \tau^2_{r0L}/\sigma^2_{dL} - 1$$

in conditional mean-squared error due to parameter estimation. This relative increase is of order $1/n$.

As in the case of linear regression, computations of relative increases in mean-squared error must be modified to some extent to study increase in error of prediction for the true essay score T_i. The conditional mean-squared error $\tau^2_{rT0L} = E(r^2_{T0L}|\mathbf{X}_i, 1 \le i \le n)$ is compared to $E([T_i - T^*_{iL}]^2) = \sigma^2_{dTL}$. Because

$$\tau^2_{rT0L} = \sigma^2_{dTL} + \tau^2_{f0L},$$

the relative increase

$$I_{TL} = \tau^2_{rT0L}/\sigma^2_{dTL} - 1$$

is $I_L\sigma^2_{dL}/\sigma^2_{dTL} > I_L$.

Similar arguments can be applied to the relative increase I_{ML} in mean-squared error for approximation of $\bar{Y}_0$ conditional on $m_0 = M > 0$. Given $m_0 = M$, the conditional mean-squared error

$$\tau^2_{rM0L} = E([\bar{Y}_0 - \hat{T}_{iL}]^2|\mathbf{X}_i, 1 \le i \le n, m_0 = M) = \tau^2_{rT0L} + \sigma^2/M$$

for predicting $\bar{Y}_0$ by $\hat{T}_{0L}$ is compared to the conditional mean-squared error σ^2_{dML} for predicting $\bar{Y}_0$ by T^*_{0L}. The relative increase in mean-squared error I_{ML} is defined as

$$I_{ML} = \tau^2_{rM0L}/\sigma^2_{dML} - 1.$$

4.3 Estimation of Mean-Squared Error

Estimation of mean-squared errors may be accomplished by use of deleted residuals; however, such a step is rather tedious for cumulative logit models if conventional software is used for analysis. An alternative approach may be based on a random division of the sample indices 1 to n into nearly equal groups (Haberman, 2006). Let $U \geq 2$ be the number of groups employed, and let each group have $[n/U]$ or $[n/U]+1$ members, where $[n/U]$ is the largest integer that does not exceed n/U. The accuracy of results is best for larger values of U, although computational convenience favors smaller U. Let K_u denote the collection of indices for group u, let n_u be the number of members of K_u, and let T^*_{iLu} be the estimate of T_i provided by applying the cumulative logit model to observations with indices i such that $1 \leq i \leq n$ and i is not in K_u. Let $r_{iLu} = \bar{Y}_i - T^*_{iLu}$ be the corresponding residual. Let

$$s^2_{rL} = n^{-1}\sum_{i=1}^{n} r^2_{iL},$$

let

$$s^2_{rL1} = (Un)^{-1}\sum_{u=1}^{U}\sum_{i=1}^{n} r^2_{iLu},$$

and let

$$s^2_{rL2} = n^{-1}\sum_{i=1}^{n} r^2_{iL*},$$

where $r_{iL*} = r_{iLu}$ for all i in K_u for $1 \leq u \leq U$. With probability 1, the conditional expectation of $n(s^2_{rL} - \sigma^2_{dL})$ given $\mathbf{X}_i$, $1 \leq i \leq n$, converges to a constant q_L. Similarly, with probability 1, the conditional expectation of $n(s^2_{rL2} - \sigma^2_{dL})$ given $\mathbf{X}_i$, $1 \leq i \leq n$, converges to $p_L U/(U-1)$, and the conditional expectation of $n(s^2_{rL1} - \sigma^2_{dL})$ given $\mathbf{X}_i$, $1 \leq i \leq n$, converges to $q_L + p_L/(U-1)$. The arguments required are very similar to those previously used with multinomial response models (Gilula & Haberman, 1994). The conditional expectation of $n(s^2_{dL} - \sigma^2_{dL})$ given $\mathbf{X}_i$, $1 \leq i \leq n$, and the conditional expectation of $n(s^2_{r0L} - \tau^2_{r0L})$ given $\mathbf{X}_i$, $1 \leq i \leq n$, both converge to 0 with probability 1. It follows that σ^2_{dL} may be estimated by

$$s^2_{dL} = U(s^2_{rL} - s^2_{rL1}) + s^2_{rL2}$$

and τ^2_{r0L} may be estimated by

$$s^2_{r0L} = s^2_{rL2} - s^2_{rL1} + s^2_{rL}.$$

In large samples, $n^{1/2}(s^2_{r0L} - \tau^2_{r0L})$ and $n^{1/2}(s^2_{dL} - \sigma^2_{dL})$ both have approximate normal distributions with mean 0 and variance equal to the variance of d^2_{iL}. Again arguments similar to those required with log-linear models can be applied (Gilula & Haberman, 1994).

4.4 Estimation of τ^2_{rT0L} and τ^2_{rM0L}

Estimations of τ^2_{rT0L} and τ^2_{rM0L} are accomplished in a manner similar to that for τ^2_{rT0} and τ^2_{rM0}. Assume that $m_i > 1$ with positive probability. Then τ^2_{rT0L} may be estimated by

$$s^2_{rT0L} = s^2_{r0L} - \hat{\sigma}^2_e$$

, and τ^2_{rM0L} may be estimated by

$$s^2_{rM0L} = s^2_{rT0L} + \hat{\sigma}^2/M.$$

4.5 Estimation of σ^2_{dTL}, σ^2_{dML} and Inflations of Mean-Squared Error

The estimate of the relative increase I_L in mean-squared error is now

$$\hat{I}_L = (s^2_{r0L} - s^2_{dL})/s^2_{dL}.$$

It also follows that an estimate of the value of I_L achieved if n^* observations are used rather than n is $\hat{I}^*_L = n\hat{I}_L/n^*$. As in the regression case, this estimate provides a guide to sample-size selection.

If $m_i > 1$ with positive probability, then σ^2_{dTL} may be estimated by

$$s^2_{dTL} = s^2_{dL} - \hat{\sigma}^2_e,$$

and I_{TL} may be estimated by

$$\hat{I}_{TL} = (s^2_{rT0L} - s^2_{dTL})/s^2_{dTL}.$$

Similar arguments apply to estimation of I_{ML}. The estimate of σ^2_{dML} is $s^2_{dML} = s^2_{dTL} + \hat{\sigma}^2/M$, so that I_{ML} has the estimate

$$\hat{I}_{ML} = (s^2_{rM0L} - s^2_{dML})/s^2_{dML}.$$

4.6 Estimation of Proportional Reduction in Mean-Squared Error Using Cross-Validation

As in the regression case, proportional reduction in mean-squared error may also be considered in terms of cross-validation. For estimation of $\bar{Y}_0$, consider the proportional reduction

$$\rho^2_{Yb0} = \frac{\tau^2_{r0C} - \tau^2_{r0L}}{\tau^2_{r0C}}$$

in mean-squared error. As in regression analysis, τ^2_{r0C} is the mean-squared error from prediction of $\bar{Y}_0$ by $\bar{Y}$. In contrast, τ^2_{r0L} is the mean-squared error from prediction of $\bar{Y}$ by $\hat{T}_{0L}$. The logical estimate of ρ^2_{Yb0} is

$$\hat{\rho}^2_{Yb0L} = \frac{s^2_{r0C} - s^2_{r0L}}{s^2_{r0C}}.$$

As the sample size n increases, $\hat{\rho}^2_{Yb0L}$ converges with probability 1 to ρ^2_{YbL} and ρ^2_{Yb0L} converges to ρ^2_{YbL}. In like manner, consider the proportional reduction

$$\rho^2_{T0L} = \frac{\tau^2_{rT0C} - \tau^2_{rT0L}}{\tau^2_{rT0C}}$$

in mean-squared error. Here τ^2_{rT0C}, as in linear regression, is the mean-squared error from prediction of T_0 by $\bar{Y}$. In contrast, τ^2_{rT0L} is the mean-squared error from prediction of T_0 by $\hat{T}_{0L}$. The corresponding estimated proportional reduction in mean-squared error is

$$\hat{\rho}^2_{T0L} = \hat{\rho}^2_{Yb0L}\frac{s^2_{r0C}}{s^2_{rT0C}},$$

and

$$\rho^2_{M0L} = \frac{\tau^2_{rM0C} - \tau^2_{rM0L}}{\tau^2_{rM0C}}$$

is estimated by

$$\hat{\rho}^2_{M0L} = \hat{\rho}^2_{Yb0L}\frac{s^2_{r0C}}{s^2_{rM0C}}.$$

4.7 A Practitioner's Guide to Cumulative Logits

Application of formulas for cumulative logit analysis is somewhat similar to that for linear regression analysis, although a few changes occur when standard software is employed. Consider once again the first prompt in the second group of essays. We will discuss results for the same three hypothetical users we considered in the practitioner's guide for the linear regression model.

First consider User 1, who is interested in the evaluation of performance in prediction of the average of the two rater scores. In this case, SAS provides the estimates $\hat{P}_{ig} - \hat{P}_{i(g-1)}$ for g from 1 to $G = 6$. Use of standard SAS functions permits computation of the estimated means

$$\hat{T}_{iL} = \sum_{g=1}^{G} g[\hat{P}_{ig} - \hat{P}_{i(g-1)}] = 1 + \sum_{g=1}^{G-1}(1 - \hat{P}_{ig}),$$

residuals

$$r_{iL} = \bar{Y}_i - \hat{T}_{iL},$$

and squared residuals r_{iL}^2. The average squared residual s_{rL}^2 is found to be 0.3139, a value a bit smaller than the corresponding regression estimate of $s_r^2 = 0.3198$. To implement cross-validation, $U = 10$ is selected. Using a series of SAS macros and computations of variables leads to an average residual for observations not used in model-fitting of $s_{rL2}^2 = 0.3330$. The average residual among all observations for all model fits with deleted data is $s_{rL1}^2 = 0.3141$. The estimated value of the conditional mean-squared error τ_{r0L}^2 for predicting a new average holistic score $\bar{Y}_0$ by the estimated cumulative logit predictor $\hat{T}_{0L}$ is then

$$s_{r0L}^2 = s_{rL2}^2 - s_{rL1}^2 + s_{rL}^2 = 0.3330 - 0.3141 + 0.3139 = 0.3328,$$

a modest improvement over the corresponding regression value of 0.3364. Similarly, one may estimate the conditional mean-squared error σ_{dL}^2 achieved by prediction of $\bar{Y}_i$ by the predictor T_{iL}^* obtained through knowledge of the joint distribution of $\bar{Y}_i$ and the features X_{ik}, $1 \leq k \leq q$. The estimate

$$s_{dL}^2 = U(s_{rL}^2 - s_{rL1}^2) + s_{rL2}^2 = 10(0.3139 - 0.3141) + 0.3330 = 0.3238$$

is a bit smaller than is s_{r0L}^2. The estimate s_{dL}^2 is also smaller than is the corresponding estimate $s_d^2 = 0.3281$ from regression analysis. The estimated proportional reduction in mean-squared error

$$\hat{\rho}_{Yb0L}^2 = \frac{s_{r0C}^2 - s_{r0L}^2}{s_{r0C}^2} = \frac{0.9496 - 0.3328}{0.9496} = 0.6513$$

is slightly larger than the corresponding value $\hat{\rho}^2_{Yb0L} = 0.6467$ from regression analysis. The estimated inflation of mean-squared error is

$$\hat{I}_L = \frac{s^2_{r0L} - s^2_{dL}}{s^2_{dL}} = \frac{0.3328 - 0.3238}{0.3328} = 0.02548,$$

slightly larger than is the corresponding value 0.02518 for regression analysis.

Next, consider User 2, who is interested in investigating prediction errors that are not due to rater variability. Here σ^2_{dTL} measures the error of prediction of the true essay score by the essay features when the joint distribution of essay features and holistic scores is known. The corresponding measure is τ^2_{rT0L} when cumulative logit parameters are estimated from the sample and the essay under study is not in the sample. As in the regression case, the estimated rater variance $\hat{\sigma}^2 = 0.1260$ is used, so that $\hat{\sigma}^2_e$, the estimated variance of the average rater error $\bar{e}_i$ for essay i, is $0.1260/2 = 0.0630$. It follows that the mean-squared error σ^2_{dTL} has estimate

$$s^2_{dTL} = s^2_{dL} - \hat{\sigma}^2_e = 0.3238 - 0.0630 = 0.2608.$$

As in regression analysis, a substantial fraction (0.0630/0.3238=0.195) of the estimated mean-squared error s^2_{dL} for predicting the average holistic score $\bar{Y}_i$ is due to rater variability. In like fashion, τ^2_{rT0L} has estimate

$$s^2_{rT0L} = s^2_{r0L} - \hat{\sigma}^2_e = 0.3321 - 0.0630 = 0.2691.$$

The difference between s^2_{rT0L} and the corresponding value $s^2_{rT0} = 0.2733$ in regression analysis reflects the difference between s^2_{r0L} and s^2_{r0}. As in regression analysis, the inflation I_{TL} in mean-squared error associated with true essay scores has estimate

$$\hat{I}_{TL} = \frac{0.2733 - 0.2608}{0.2608} = 0.03164$$

that is larger than $\hat{I}_L$, the corresponding estimated inflation in mean-squared error for predicting observed average essay scores. The values of $\hat{I}_{TL}$ for cumulative logit analysis and $\hat{I}_T = 0.03117$ for regression analysis are quite similar.

As in regression analysis, the proportional reduction ρ^2_{TL} in mean-squared error for predicting true essay score by essay features when all joint distributions are known is larger than the corresponding proportional reduction ρ^2_{YbL} in mean-squared error for predicting average essay score by essay features. When sampling is required, the proportional reduction ρ^2_{T0L} in mean-squared error for predicting true essay score by essay features is normally smaller than ρ^2_{TL}. The estimated proportional reduction in mean-squared error is

$$\hat{\rho}^2_{T0L} = (0.8892 - 0.2691)/0.8892 = 0.6974,$$

a value slightly higher than the corresponding estimate $\hat{\rho}^2_{T0} = 0.6926$ from regression analysis.

Finally, consider User 3, who is interested in predicting model performance for predicting the performance of a single rater ($M = 1$) using the data based on two raters per essay. Coefficients σ^2_{dML}, τ^2_{rM0L}, I_{ML}, and ρ^2_{YbML} are considered here. For the example, the estimated mean-squared error s^2_{dML} for known population characteristics is

$$s^2_{dTL} + \hat{\sigma}^2 = 0.2608 + 0.1260 = 0.3868,$$

while the estimated cross-validation mean-squared error is

$$s^2_{rM0L} = s^2_{rT0L} + \hat{\sigma}^2 = 0.3951.$$

The estimated proportional reduction in mean-squared error is

$$\hat{\rho}^2_{M0L} = (1.015 - 0.3951)/1.015 = 0.6108,$$

a value slightly lower than the corresponding regression estimate $\hat{\rho}^2_{M0} = 0.6066$. Inflation of mean-squared error is similar to that found in the regression case. For cumulative logits,

$$\hat{I}_{ML} = (0.3951 - 0.3868)/0.3951 = 0.02133.$$

The comparable value for regression analysis is $\hat{I}_M = 0.02113$.

4.8 Results From Analysis of the Data

Data analysis with cumulative logits was performed in a manner similar to that for regression. We used $U = 10$ in the cross-validation calculations. First, each essay in

Table 8.
Summary of Cumulative Logit Analysis of Individual Essay Prompts Within Groups

	Group 1		Group 2		Group 3		Group 4	
Statistic	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	S.D.
n	495.9	6.5	456.0	62.0	467.0	14.4	499.8	6.2
s^2_{r0L}	0.335	0.061	0.350	0.118	0.230	0.039	0.443	0.184
s^2_{r0C}	0.871	0.084	1.234	0.255	1.688	0.322	1.215	0.297
$\hat{I}_L$	0.0198	0.0034	0.0227	0.0055	0.0233	0.0036	0.0195	0.0040
$n\hat{I}_L$	9.60	1.67	10.13	1.83	10.86	1.61	9.74	1.97
$\hat{\rho}^2_{Yb0L}$	0.613	0.079	0.699	0.137	0.859	0.033	0.643	0.096
s^2_{rT0L}			0.292	0.125	0.115	0.036	0.307	0.187
s^2_{rT0C}			1.175	0.261	1.572	0.323	1.080	0.301
$\hat{I}_{TL}$			0.0290	0.0091	0.0504	0.0120	0.0335	0.0123
$n\hat{I}_{TL}$			12.90	3.16	23.6	5.9	16.8	6.1
$\hat{\rho}^2_{T0L}$			0.736	0.144	0.926	0.019	0.732	0.120
s^2_{r10L}			0.408	0.116	0.345	0.046	0.570	0.178
s^2_{r10C}			1.292	0.251	1.803	0.322	1.351	0.293
$\hat{I}_{1L}$			0.0191	0.0050	0.0153	0.0026	0.0144	0.0034
$n\hat{I}_{1L}$			8.55	1.78	7.15	1.15	7.17	1.64
$\hat{\rho}^2_{10L}$			0.667	0.134	0.802	0.046	0.574	0.085

the four groups of prompts was analyzed. A summary of results is reported in Table 8. Several basic conclusions appear possible for these groups of prompts. Cumulative logit analysis results in a notable reduction in mean-squared error compared to linear regression. Inflation in mean-squared error due to estimation of parameters is quite comparable to that encountered with regression. Thus sample-size recommendations are similar to those for regression analysis. The general pattern of relative success of prediction for different groups, at least as measured by proportional reduction in mean-squared error, is the same as for regression analysis.

4.9 Combining Essays in Groups

As in the regression case, an alternative approach summarizes the data by looking at the prediction of a score for a group of essays. The initial approach is to use distinct

Table 9.
Summary of Cumulative Logit Analysis of Essay Prompts Within Groups:Distinct Coefficients for Each Essay

Statistic	Group 1	Group 2	Group 3	Group 4
n	36,697	6,384	12,143	7,997
s^2_{r0}	0.352	0.366	0.288	0.443
s^2_{r0C}	0.892	1.320	1.698	1.355
$\hat{I}$	0.0194	0.0225	0.0232	0.0196
$n\hat{I}$	711.15	143.57	281.71	156.78
$\hat{\rho}^2_{Yb0}$	0.605	0.723	0.856	0.636
s^2_{rT0}	0.109	0.292	0.115	0.307
s^2_{rT0C}	0.668	1.262	1.582	1.219
$\hat{I}_T$	0.0620	0.0271	0.0476	0.0285
$n\hat{I}_T$	2,273.86	172.88	578.40	228.12
$\hat{\rho}^2_{T0}$	0.837	0.769	0.928	0.748
s^2_{r10}	0.347	0.408	0.345	0.578
s^2_{r10C}	0.906	1.378	1.813	1.491
$\hat{I}_1$	0.0186	0.0192	0.0153	0.0149
$n\hat{I}_1$	683.28	122.78	186.20	119.43
$\hat{\rho}^2_{10}$	0.617	0.704	0.810	0.612

parameters for each prompt, so that K prompts involve $K(q+G)$ parameters. As evident from Table 9, results for this approach are quite similar to those shown in Table 8, except that the proportional reduction in mean-squared error is increased slightly because it is computed relative to a constant predictor for all essays in the entire family rather than relative to a constant predictor for each prompt. The number of prompts scored in Group 1 is sufficient for analysis related to true scores; however, results for true scores and for exactly one score should be approached with caution given that the sampling assumptions appear questionable. The increase in the estimated product of sample size by relative inflation of mean-squared error primarily reflects the increased number of predictors present in the analysis.

As with a previous regression analysis, one may consider a model in which, for all prompts in a group, the slope for a feature is constant but the intercept is sum of a score effect and a prompt effect, so that $q+G+K-1$ parameters are used. Results are

Table 10.

Summary of Cumulative Logit Analysis of Essay Prompts Within Groups: Distinct Intercepts for Each Essay, Common Slopes

Statistic	Group 1	Group 2	Group 3	Group 4
n	36,697	6,384	12,143	7,997
s^2_{r0L}	0.342	0.361	0.231	0.454
s^2_{r0C}	0.892	1.320	1.698	1.355
$\hat{I}_L$	0.0023	0.0034	0.0036	0.0027
$n\hat{I}_L$	83.62	21.49	42.98	21.64
$\hat{\rho}^2_{Yb0L}$	0.617	0.726	0.864	0.665
s^2_{rT0L}	0.117	0.303	0.115	0.318
s^2_{rT0C}	0.668	1.262	1.582	1.219
$\hat{I}_{TL}$	0.0067	0.0040	0.0071	0.0039
$n\hat{I}_{TL}$	245.53	25.62	86.27	30.91
$\hat{\rho}^2_{T0L}$	0.825	0.760	0.927	0.739
s^2_{r10L}	0.355	0.4194	0.346	0.590
s^2_{r10C}	0.906	1.378	1.813	1.491
$\hat{I}_{1L}$	0.0022	0.0029	0.0024	0.0021
$n\hat{I}_{1L}$	80.47	18.51	28.62	16.65
$\hat{\rho}^2_{10L}$	0.608	0.696	0.809	0.604

summarized in Table 10. Relative to using distinct intercepts and regression slopes for each prompt, estimated losses in mean-squared error are very limited (Groups 1, 2, and 4) or nonexistent (Group 3). This approach has much more modest sample-size requirements than does the approach with individual regression coefficients for each prompt and produce a tolerable inflation of mean-squared error with about a tenth of the essays for each prompt. Note that this approach does require that the group contains a substantial number of prompts.

The simplest model for a group of essays ignores the prompt entirely, so that only $G + q$ parameters are needed. Results are summarized in Table 11. As in the regression case, although the inflations of mean-squared error are very small, the tradeoff is a substantial increase in the actual mean-squared error in Group 1 and in Group 2. Losses in mean-squared error are also encountered in the other groups, but they are very small in Group 3 and modest in Group 4. As in the regression case, the virtue of the approach with

Table 11.
Summary of Cumulative Logit Analysis of Essay Prompts Within Groups: Common Intercepts and Common Regression Slopes

Statistic	Group 1	Group 2	Group 3	Group 4
n	36,697	6,384	12,143	7,997
s^2_{r0L}	0.384	0.425	0.235	0.482
s^2_{r0C}	0.892	1.320	1.698	1.355
$\hat{I}_L$	0.0002	0.0011	0.0009	0.0010
$n\hat{I}_L$	9.13	6.89	10.94	7.96
$\hat{\rho}^2_{Yb0L}$	0.569	0.678	0.862	0.644
s^2_{rT0L}	0.160	0.367	0.120	0.347
s^2_{rT0C}	0.668	1.262	1.582	1.219
$\hat{I}_{TL}$	0.0006	0.0013	0.0018	0.0014
$n\hat{I}_T$	21.96	7.99	21.50	11.08
$\hat{\rho}^2_{T0}$	0.760	0.709	0.924	0.716
s^2_{r10}	0.398	0.483	0.350	0.618
s^2_{r10C}	0.906	1.378	1.813	1.491
$\hat{I}_1$	0.0002	0.0010	0.0006	0.0008
$n\hat{I}_1$	8.82	6.07	7.34	6.21
$\hat{\rho}^2_{10}$	0.569	0.650	0.807	0.585

a common equation for each prompt is that the sample-size requirements for the group are similar to those for a single prompt. Thus one could consider use of several hundred essays for a complete group.

Each cumulative logit model performs better than its corresponding regression model, making the cumulative logit approach attractive. The gains for the cumulative logit method are somewhat variable.

5. Conclusions

This paper employs cross-validation methods to assess sample-size requirements both for cumulative logit and ordinary regression models. Sample-size requirements depend on the application and on the e-rater features used. In typical cases in which content analysis is not employed and the only object is to score individual essays to provide feedback to the examinee, it appears that several hundred essays are quite sufficient to limit variance inflation to less than 5%. For a large family of essays, fewer than 100 essays per prompt may often be adequate.

Sample-size requirements when content is assessed appear to be much larger (Haberman, 2006), and proper analysis requires significant modifications to currently used software.

These recommendations are not appropriate for all potential uses. If e-rater is used within an equated assessment or if substantial groups of students are to be compared by using e-rater, then sample-size requirements may be much higher.

For the examples in this report, using common parameters for essay features for all prompts within a family appears to be an attractive option, but completely ignoring all effects related to prompts appears to be less attractive. Nonetheless, for the third group of prompts, ignoring all prompt effects was strikingly successful. Treatment of groups of prompts therefore appears to require treatment on a case-by-case basis.

An important finding in this paper is that the cumulative logit model typically performed somewhat better than did ordinary regression analysis. Although cumulative logit analysis requires more difficult cross-validation than does ordinary regression analysis, the cross-validation is hardly burdensome using standard statistical software. For electronic essay scoring, cumulative logit analysis should be considered a very attractive alternative to regression analysis. Conceptually, a cumulative logit model makes more sense than an ordinary regression model in electronic essay scoring because the observed responses (i.e., the essay scores) are categorical, with usually four to six categories. For all the groups of essays examined, the average mean-squared error for the cumulative logit model is less, often substantially less, than that for the ordinary regression model. In addition, the requirements in terms of sample size for the cumulative logit model appear to be comparable to those for ordinary regression. Consequently, our research indicates that it may be worthwhile to replace the ordinary regression model with a cumulative logit model in electronic essay-scoring software packages.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v.2. *Journal of Technology, Learning and Assessment*, *4*(3), 1-29.

Attali, Y., Burstein, J., & Andreyev, S. (2003). *E-rater® version 2.0: Combining writing analysis feedback with automated essay scoring.* (Unpublished manuscript)

Bock, R. D. (1973). *Multivariate statistical methods in behavioral research.* New York: McGraw-Hill.

Box, G. E. P. (1954). Some theorems on quadratic form applied to the study of analysis of variance problems. *Annals of Mathematical Statistics*, *25*, 290-302.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *AI Magazine*, *25*(3), 27-36.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: John Wiley.

Feng, X., Dorans, N. J., Patsula, L. N., & Kaplan, B. (2003). *Improving the statistical aspects of e-rater®: Exploring alternative feature reduction and combination rules* (Research Rep. No. RR-03-15). Princeton, NJ: ETS.

Gilula, Z., & Haberman, S. J. (1994). Models for analyzing categorical panel data. *Journal of the American Statistical Association*, *89*, 645–656.

Haberman, S. J. (1989). Concavity and estimation. *The Annals of Statistics*, *17*, 1631–1661.

Haberman, S. J. (2006). Electronic essay grading. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics* (Vol. 26, pp. 205–233). Amsterdam: North-Holland.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

McCullagh, P., & Nelder, J. A. (1989). *Generalized linear models* (2nd ed.). Boca Raton, FL: Chapman and Hall.

Neter, J., Kutner, M. H., Nachtsheim, C. J., & Wasserman, W. (1996). *Applied linear regression models.* Homewood, IL: Irwin.

Pratt, J. W. (1981). Concavity of the log likelihood. *Journal of the American Statistical*

Association, *76*, 103–109.

Rao, C. R., & Mitra, S. K. (1971). *Generalized inverse of matrices and its applications.* New York: John Wiley.

Weisberg, S. (1985). *Applied linear regression* (2nd ed.). New York: John Wiley.

Appendix

Proof That an Estimate of σ_d^2 is $s_d^2 = (s_r^2 + s_{r0}^2)/2$

Equations 5 and 6 suggest that for large n,

$$n(\tau_{r0}^2 - \sigma_d^2) \approx \sigma_d^2 + Z, \tag{A1}$$

where $Z = \operatorname{tr}([\operatorname{Cov}(\mathbf{X})]^{-1} \operatorname{Cov}(d\mathbf{X}))$. Equation A1 suggests that

$$\begin{aligned}
\sigma_d^2 &\approx (\tau_{r0}^2 - Z/n)(1 + 1/n)^{-1} \\
&\approx (\tau_{r0}^2 - Z/n)(1 - 1/n) \\
&\approx \tau_{r0}^2 - \tau_{r0}^2/n - Z/n.
\end{aligned}$$

By Equation 7,

$$\sigma_d^2 \approx \tau_{r0}^2 - \tau_{r0}^2/n - \frac{1}{2}[\tau_{r0}^2 - s_r^2(1 + 2/n)]\cdot$$

Replacing τ_{r0}^2 by its estimate s_{r0}^2,

$$\begin{aligned}
\sigma_d^2 &\approx \frac{s_{r0}^2 + s_r^2}{2} + (s_r^2 - \tau_{r0}^2)/n \\
&\approx \frac{s_{r0}^2 + s_r^2}{2}.
\end{aligned}$$